

Evolution of sparsity and modularity in a model of protein allostery

Mathieu Hemery and Olivier Rivoire
*CNRS, LIPhy, F-38000 Grenoble, France and
 Univ. Grenoble Alpes, LIPhy, F-38000 Grenoble, France*

The sequence of a protein is not only constrained by its physical and biochemical properties under current selection, but also by features of its past evolutionary history. Understanding the extent and the form that these evolutionary constraints may take is important to interpret the information in protein sequences. To study this problem, we introduce a simple but physical model of protein evolution where selection targets allostery, the functional coupling of distal sites on protein surfaces. This model shows how the geometrical organization of couplings between amino acids within a protein structure can depend crucially on its evolutionary history. In particular, two scenarios are found to generate a spatial concentration of functional constraints: high mutation rates and fluctuating selective pressures. This second scenario offers a plausible explanation for the high tolerance of natural proteins to mutations and for the spatial organization of their least tolerant amino acids, as revealed by sequence analyses and mutagenesis experiments. It also implies a faculty to adapt to new selective pressures that is consistent with observations. Besides, the model illustrates how several independent functional modules may emerge within a same protein structure, depending on the nature of past environmental fluctuations. Our model thus relates the evolutionary history and evolutionary potential of proteins to the geometry of their functional constraints, with implications for decoding and engineering protein sequences.

PACS numbers:

Proteins are well known to be highly tolerant to mutations [1]. Statistical analyses of protein sequences [2] and saturated mutagenesis experiments [3] are now revealing the spatial architecture of this robustness: in several proteins, the amino acids most essential to the function are organized in small, structurally connected clusters of interacting and coevolving residues, called protein sectors [4]. For instance, in PDZ domains, a family of small interaction domains, a sector connects the ligand binding pocket to an opposite surface site [2, 3], which regulates allosterically the active site in at least one member of the family [5]. Similar sectors have been found and experimentally investigated in other protein families, which also consist of small structurally connected subsets of residues and, in many cases, mediate allostery [6, 7, 8]. Furthermore, several quasi-independent sectors have been found to co-exist within a same protein domain [4].

This spatial concentration of functional constraints within a protein structure is presently unexplained. It may be inherent to the physical properties of proteins, including the functional properties for which they were selected. For instance, when the function involves binding to a ligand, the residues structurally closer to the ligand may be expected to be functionally more important. We shall show, however, that such structural heterogeneities are not needed to explain a spatial concentration of functional constraints within a protein structure. To this end, we introduce below a simple mathematical model in which all "residues" are *a priori* equivalent, but where a sparse sector can nevertheless arise as a consequence of fluctuations during the evolutionary process.

The role of evolutionary history in shaping biological organizations has been discussed previously in relation to

modularity, the generic decomposition of biological networks into subnetworks [9], of which the presence of several independent sectors within a same protein structure is an illustration [4]. Explanations for the origin of modularity broadly fall in two classes [10]: first, those based on the combinatorial properties of the process generating new variations, e.g., gene duplications and recombinations [11], and, second, those invoking the history of selective pressures, notably the particular structure of environmental fluctuations [12]. In proteins, combinatorial reassortment may thus explain multi-domain architectures, and selection the decomposition of a domain into sectors that are intermingled along the sequence [4]. In general, however, the variational and selective factors are non-exclusive, and may contribute jointly to the emergence of modules [13]. Besides the question of their origin, the implications of modular architectures for future evolution have also been extensively studied in terms of resilience to mutations, or "robustness", and in terms of faculty to adapt, or "evolvability" [14], two properties found to have a complex relationship [15, 16, 17].

The presence of a single sector in the network of interacting residues forming the structure of a protein corresponds to a degenerate form of modularity, better referred to as "sparsity" [18]. Here, we demonstrate in the context of a physical model of protein evolution how sparsity generically emerges in the form of a spatial concentration of functional constraints from fluctuations during the evolutionary process. These fluctuations may involve variational or selective factors, and may promote robustness and/or evolvability to varying degrees. The phenomenon that we describe is more elementary than the evolution of modularity, which arises when the fluctua-

tions have some additional structure, in relation to the structure of the function itself.

A model for the evolution of allostery

To illustrate the role of evolutionary history in a context where structural heterogeneities are minimized, we introduce a model defined on a regular structure, and consider an allosteric property, which may in principle involve the entire structure. In the spirit of previous theoretical studies of protein evolution [19], we present this model in the generic framework of spin glasses [20], but the Gaussian spin glass [21] that we analyze more specifically is also closely linked to models of elastic networks [22].

We may derive our model starting from a general expression for the energy of a protein,

$$E = - \sum_i K_0(a_i, \sigma_i, \varepsilon(r_i)) - \sum_i K_1(a_i, a_{i+1}, \sigma_i, \sigma_{i+1}) - \sum_{i,j} K_2(a_i, a_j, r_i, r_j, \sigma_i, \sigma_j), \quad (1)$$

where a_i indicates the amino acid at position i along the chain, r_i its mean position, say of its alpha carbon, and σ_i its physical state, say its orientation and fluctuations around the mean position. The term $K_0(a_i, \sigma_i, \varepsilon(r_i))$ represents an interaction energy between the amino acid a_i and its local environment $\varepsilon(r_i)$, which may include the solvent and/or the amino acids of another protein, $K_1(a_i, a_{i+1}, \sigma_i, \sigma_{i+1})$ represents the bonding energy between successive amino acids along the chain, and $K_2(a_i, a_j, r_i, r_j, \sigma_i, \sigma_j)$ the interactions of residues far apart along the chain but brought together upon folding.

While others have studied the incidence of evolutionary parameters on protein structure and stability [23, 24], we focus here on the evolution of amino-acid specific variables in the context of a fixed structure, to analyze the structural organization of functional constraints in members of a protein family sharing a common fold. We thus fix the r_i at the nodes of a lattice, where only nearest neighbors have non-zero interactions. For simplicity, and to minimize structural heterogeneities, we also ignore the distinction between bond and non-bond energies, so that

$$E = - \sum_{\langle i,j \rangle} K(a_i, a_j, \sigma_i, \sigma_j) - \sum_i K_0(a_i, \sigma_i, \varepsilon(r_i)), \quad (2)$$

where $\langle i, j \rangle$ indicates neighboring sites on the lattice.

We further assume that the σ_i take real values and that K has the form $K(a_i, a_j, \sigma_i, \sigma_j) = J(a_i, a_j)\sigma_i\sigma_j$. Similarly, we assume that the environment around i is represented by a real number h_i with K_0 of the form $K_0(a_i, \sigma_i, \varepsilon(r_i)) = h_i\sigma_i$ (more generally, h_i could depend

on a_i). We thus arrive at an energy of the form

$$E(\sigma|a, h) = - \sum_{\langle i,j \rangle} J(a_i, a_j)\sigma_i\sigma_j - \sum_i h_i\sigma_i, \quad (3)$$

which is formally the energy of a spin glass [25], where spins σ_i interact in the context of given (quenched) couplings $J(a_i, a_j)$ and fields h_i .

These simplifications are drastic but retain the essential relationships between the variables of the problems: the amino acid a_i , which may be subject to evolution, the environmental variables h_i , which may vary with time, and the physical variables σ_i , which are subject to short-range interactions constrained by the overall structure and are dependent on the identity of the amino acids and on the environment.

In this framework, we can apply the standard approach of statistical mechanics and sum over the internal degrees freedom σ_i to compute from $E(\sigma|a, h)$ a free-energy $F(a, h)$. We can thus define a free energy of binding with an external ligand: the presence of a ligand corresponds indeed to a field h' differing from the field h in its absence, so a binding free energy is obtained as the difference $F(a, h') - F(a, h)$. On the other hand, mutations induce a difference of free energy of the form $F(a', h) - F(a, h)$.

Here, we consider a cylindrical lattice where two different ligands can bind at the two opposite open ends (Fig. 1A). This allows us to quantify the preferential binding of one of the ligands in presence of the other, which corresponds to an allosteric regulation [26]. Following the terminology used for allosteric proteins, we call "regulatory site" (abbreviated in 'reg') the upper end of the cylinder, "modulator" the ligand binding to it, "active site" ('act') its lower end, and "endogenous ligand" the ligand binding to it. Taking the interaction of site i with the solvent to correspond to $h_i = 0$, we thus have

$$E(\sigma|a, h) = - \sum_{\langle i,j \rangle} J(a_i, a_j)\sigma_i\sigma_j - \sum_{i \in \text{reg}} h_i^{\text{reg}}\sigma_i - \sum_{i \in \text{act}} h_i^{\text{act}}\sigma_i, \quad (4)$$

where $h_i^{\text{reg}} = m_i$ in the presence of a modulator characterized by the vector m_i ($i \in \text{reg}$), $h_i^{\text{reg}} = 0$ in its absence, and $h_i^{\text{act}} = \ell_i$ in the presence of an endogenous ligand characterized by the vector ℓ_i ($i \in \text{act}$), $h_i^{\text{act}} = 0$ in its absence.

We define allostery as a more favorable interaction with a ligand ℓ in the presence of a modulator m . It is quantified thermodynamically in terms of free energy differences [27], as

$$\phi(J|m, \ell) = \Delta F(\ell|0) - \Delta F(\ell|m) \quad (5)$$

where $\Delta F(\ell|0) \equiv F(\ell|0) - F(0|0)$ represents the binding free energy of ℓ in absence of m , and $\Delta F(\ell|m) \equiv F(\ell|m) - F(0|m)$ in its presence, as illustrated in Fig. 1A (the amino acids are fixed in these expressions).

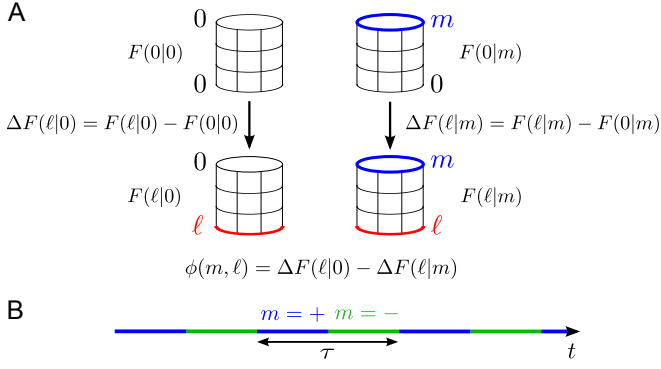


FIG. 1: **A.** A model of protein allostery is defined as a spin glass on a cylindrical lattice. In this model, spins σ_i are on the nodes to represent the physical states of residues, and couplings J_{ij} are on the edges to represent interactions between residues, subject to evolution. Interactions with a modulator m and/or a ligand ℓ are modeled by fields h_i applied to the sites i at the open ends of the cylinder. Allostery is quantified by $\phi(m, \ell)$, the difference between the binding free energy of ℓ in the presence of m , $\Delta F(\ell|m)$, and in its absence, $\Delta F(\ell|0)$. **B.** The sequences of the modulator and ligand are defined by the values and signs of the fields h_i , with $h_i = 0$ representing an interaction with the solvent. Evolution is performed over a population of systems with selection for allosteric efficiency and with mutations affecting the couplings J_{ij} at a rate μ . A given generation is selected for allostery with given ℓ, m but when the environment fluctuates, different generations may experience different ℓ, m . By symmetry, varying ℓ or m is equivalent and we fix the sequence of the ligand to $h_i = +1$ for all i at the bottom of the cylinder when ℓ is present, and vary only the sequence of the modulator every $\tau/2$ generations, between $h_i = +1$ for i at the top ($m = +$) and $h_i = -1$ ($m = -$). Modulators or ligands with non-uniform sequences can be also considered as in Fig. S6.

We simulate an evolutionary dynamics by a standard genetic algorithm [28], whereby a population of P systems undergoes repeated cycles of selection, reproduction and mutation. Selection and reproduction are based on allosteric efficiency, as defined by Eq. (5), with systems with larger $\phi(J)$ generating more offsprings. Mutations of the amino acids correspond to changes of the couplings; for simplicity, instead of introducing an arbitrary matrix $J(a, b)$, we assume that a mutation randomly change the value of a single coupling $J_{ij} = J(a_i, a_j)$ at a rate μ per generation, independently of the other couplings (Materials and methods); we verified, however, that explicitly mutating amino acids at the level of sites, which affect simultaneously several couplings, lead to similar results.

Numerical simulations of evolutionary dynamics are generally limited by the computational cost of estimating the fitness of each individual. Here, four free energies of spin glass models on finite-dimensional lattices are involved; their computations would be very demanding if considering Ising spins $\sigma_i = \pm 1$ [29], but by considering a Gaussian model [21], for which $\phi(J)$ can be expressed an-

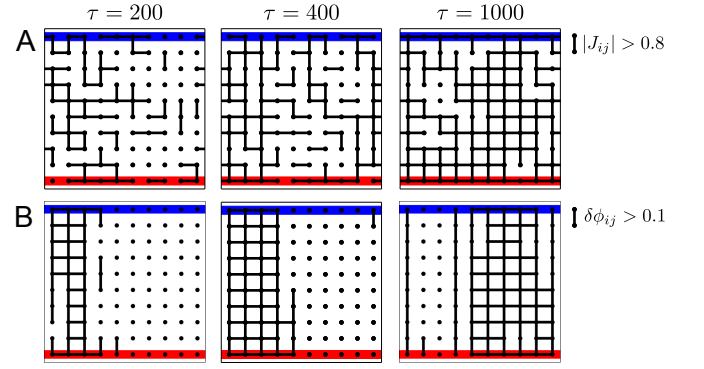


FIG. 2: Examples of systems obtained from an evolutionary dynamics with mutation rate $\mu = 5.10^{-5}$ and different periods τ of fluctuations of selective pressure ($\tau = 200, 400, 1000$). **A.** Couplings J_{ij} with large absolute values, $|J_{ij}| > 0.8$. **B.** Couplings J_{ij} inducing a large loss in allosteric efficiency when mutated, $\delta\phi_{ij} > 0.1$ (see Fig. S1 for other values of the cut-offs). A third approach, based on coevolution, can also reveal the same concentration of functional constraints (Fig. S3). The figures display the fittest individual in a population of $P = 500$ individuals prior to a change of environment.

alytically for any couplings J_{ij} and any geometry of the lattice, the computations are reduced to the inversion of a matrix (Materials and methods). In this Gaussian model, the spins σ_i take arbitrary real values, but the couplings J_{ij} need to be bounded: we thus mutate the couplings by drawing them uniformly in $[-1, +1]$. This Gaussian model may also be viewed as an elastic network model [22] with a single degree of freedom per site and non-uniform "spring constants".

Evolutionary concentration of functional constraints

The outcome of the evolutionary dynamics is contingent on the series of modulator and ligand sequences that the successive generations encounter (Fig. 1B). When these sequences are constant over time, say $m = (+1, \dots, +1)$ and $\ell = (+1, \dots, +1)$ at all time, systems evolve maximal couplings $|J_{ij}| \simeq 1$ at all sites. This implementation of the couplings optimizes the allosteric efficiency ϕ and epitomizes an absence of sparsity. Repeating the same simulations with a modulator that alternates with period τ between two sequences, $m = (+1, \dots, +1)$ and $m = (-1, \dots, -1)$, yields a qualitatively different outcome: the smaller τ is, the fewer are the large couplings, as illustrated in Fig. 2A.

Allostery requires strong couplings, but not all strong couplings need to be functionally significant: if a strong coupling is defined by $|J_{ij}| > 0.8$ as in Fig. 2A, we may indeed expect $\sim 20\%$ of strong couplings even in absence of any selection, only because the J_{ij} are mutated to random values in $[-1, 1]$ (0.8 is an arbitrary cut-off but

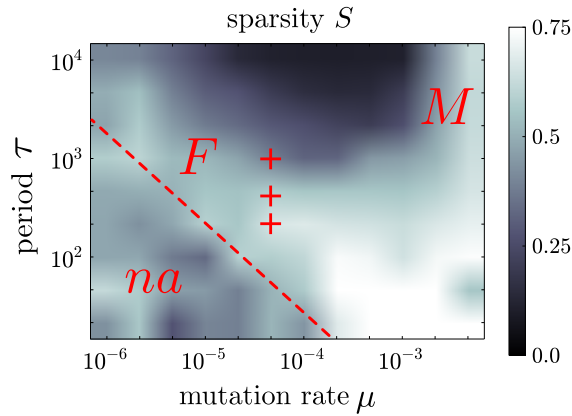


FIG. 3: Sparsity of evolved proteins as a function of the mutation rate μ and the period τ of fluctuating selective pressures. Sparsity is defined as the fraction of couplings with $\delta\phi_{ij} < 0.1$ (non-represented couplings in Fig. 2B). Below the dotted line, the environmental fluctuations are too fast for the population to follow them, and the systems are non-adapted (region *na*, see also Fig. S5). Sparse systems are found in two ranges of parameters: for intermediate values of $\mu\tau$, where they are driven by fluctuating selective pressures (region *F*, including the three systems of Fig. 2 indicated by crosses), and for high values of μ , where they are driven by a large mutational load (region *M*).

other values lead to a similar conclusion, see Fig. S1). As a more relevant measure of "functional significance", we may consider instead the "fitness value" $\delta\phi_{ij}$ of J_{ij} , defined as the maximal cost that a mutation of J_{ij} can cause to the allosteric efficiency ϕ (Material and methods). We thus identify functionally significant couplings J_{ij} by $\delta\phi_{ij} > \epsilon$, with for instance $\epsilon = 0.1$ in Fig. 2B (see Fig. S1 for other values of ϵ). This criterion, closer to what has been experimentally measured [3], reveals distinctly the presence of a connected subset of functional couplings joining the regulatory and active sites (Fig. 2B).

This subset of functionally significant couplings, which breaks the rotational invariance of the cylinder and whose location varies from simulation to simulation, displays several features reminiscent of protein sectors observed in natural proteins [2, 4]: (i) it is overall structurally connected (Fig. 2B); (ii) it has a hierarchical organization: less significant couplings are peripheral to more significant ones, as shown by varying the value of the cut-off ϵ defining functional significance (Fig. S1); (iii) it is evolutionarily conserved: its location is stable over multiple periods along a given evolutionary trajectory (Fig. S2); (iv) its couplings are coevolving, as shown by a statistical analysis of a 'multiple sequence alignment' obtained from independent evolutionary trajectories with a common origin (Fig. S3).

As indicated by Fig. 2B, the smaller the period τ of the fluctuations in selective pressure, the smaller the sec-

tor. The temporal structure of past selective objectives is thus encoded geometrically in the couplings. More precisely, we may define the sparsity of a system as the fraction S of its couplings J_{ij} with $\delta\phi_{ij} < 0.1$ (the fraction of non-represented couplings in Fig. 2B). This measure of sparsity, is represented in Fig. 3 as a function of the mutation rate μ and of the period τ . At not too high mutation rates (see below), it scales with $\mu\tau$, the number of mutations per period; more precisely, it scales with $\mu\tau P$, the total number of mutations in a population of size P (Fig. S4).

While sparsity arises at the expense of instantaneous fitness, here defined by the allosteric efficiency ϕ (Fig. S5), it favors the "evolvability" [15] of the population, which can be quantified as the fraction of random mutations conferring a noticeable fitness advantage following a change of selective objective (Suppl. Appendix, Figs. S9C-S10C). The evolution of sparsity also implies an increased mutational "robustness" [30], defined as the fraction of mutations that do not affect noticeably the fitness (Suppl. Appendix, Figs. S9B-S10B).

The period τ is not the only feature of the environmental fluctuations that affects the size of a sector: so does the diversity of these fluctuations. For a given τ , the sparsity thus decreases linearly with the sequence similarity between the two alternating modulators (Fig. S6A). But while the similarity between successive modulators is determining, their exact sequence is not: replacing the sequences $m = (+1, \dots, +1)$ and $m = (-1, \dots, -1)$ by arbitrary sequences of ± 1 , or even imposing new randomly chosen modulators at each period, does not affect significantly the outcome (Fig. S6B). This observation illustrates a capacity of "generalization" [31]: the sparse systems, which are more prompt to re-adapt to a modulator previously encountered in their history, are as prompt to adapt to a modulator never encountered.

Another factor besides the fluctuations of selective pressures can induce a sector: a large mutational load. While for small mutations rates μ the sparsity is controlled by the dimension-less parameters $\mu\tau$, for large mutation rates it is controlled by μ nearly independently of τ (Figs. 3 and S10A). The critical value of the mutation rate, $\mu_c \sim N^{-1}$, corresponds to the "error-threshold" for a system of size N (here the number of links ij), i.e., to the maximal mutation rate at which a system of this size can faithfully replicate [32]. For $\mu > \mu_c$, the systems thus evolve a sector of size $\sim (\mu N)^{-1}$, which is the largest size allowed by the mutational load. The sparse systems obtained at high μ are more robust and less evolvable than the sparse systems obtained in fluctuating environments at low μ , but nevertheless more evolvable than the non-sparse systems obtained at lower values of μ (Figs. S9-S10). A difference is also apparent at the population level: the variance in the population of the couplings outside the sectors is low at low μ and large at large μ .

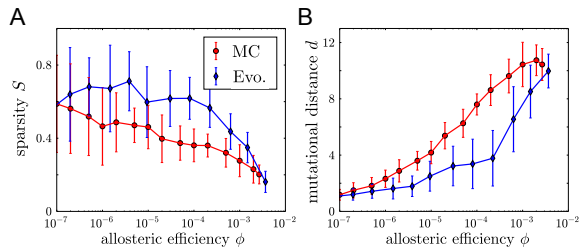
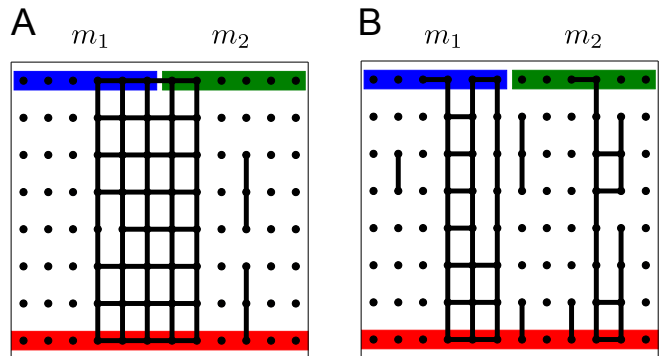


FIG. 4: **A.** Sparsity as a function of allostery efficiency (fitness) for evolved systems (in blue, various $\mu < 10^{-2}$ and $\tau > 10$) and for typical systems with same fitness (in red), obtained by Monte Carlo sampling (Materials and methods). **B.** Distance to a system with different function (allostery induced by a modulator m' different from the one m for which the system was last selected), measured by the minimal number of beneficial mutations needed to reach an equivalent allostery efficiency after the change $m \rightarrow m'$, for evolved systems (blue) and typical systems (red). Systems evolved in a fluctuating environment are thus atypical amongst systems with equivalent fitness value for being sparser and closer to solutions to new selective challenges.

Localization in sequence space

Functional proteins represent only a tenuous subset of all potential proteins [33]. In our model, we find that within this subset, proteins with a sparse sector are themselves rare: typical systems with a given fitness ϕ are significantly less sparse than systems with the same fitness but resulting from an evolution in fluctuating environments (Fig. 4A). This observation implies that sparsity in the evolved systems is not just a consequence of the fitness being curbed by the environmental fluctuations. The sparse systems are, besides, not distributed randomly in sequence space, but localized in evolvable regions of this space: they are at shorter mutational distance to solutions to alternative selective pressures (Fig. 4B). This phenomenon of localization is generic and has been illustrated previously in the context of fitness landscapes defined on small, schematic sequence spaces [34, 35].

Our model, however, displays two features absent from simpler models. First, it relates the topology of the fitness landscape, defined in sequence space, to the geometry of the functional constraints, defined in real space: gradients in fitness thus map to sector positions, where adaptive mutations occur, while plateaux in fitness map to non-sector positions, where neutral mutations occur. Second, as typical to high-dimensional spaces, the results are partly non-intuitive: a system localized between two alternating fitness peaks is thus ipso facto localized near a large family of related fitness peaks (Fig. 4B), a feature that underlies the faculty of generalization [31], or "promiscuity" [36], previously noted.



Proba. of modularity:

	$\tau = 100$	$\tau = 200$
m_1, m_2 vary together	0.03	0.10
m_1, m_2 vary one at a time	0.82	0.51

FIG. 5: Typical outcomes for a variant of the model where the regulatory site is partitioned into two sub-sites, each associated with an independently varying modulator, and where selection is on allostery in the presence of at least one of the two modulators m_1 and/or m_2 . **A.** A non-modular system, with a single sector localized at the interface between the two regulatory sites. **B.** A modular system, with two distinct sectors. These two systems are the (stable) outcomes of two distinct evolutionary trajectories with same evolutionary parameters ($\tau = 100$, $\mu = 5 \cdot 10^{-4}$). As for the location of the sector in Fig. 2, the difference stems from stochastic effects. The table indicates the probability to obtain a modular system for two values of the period τ and for m_1, m_2 varying either simultaneously or consecutively.

Modularity

The concentration of functional constraints may take different geometrical forms depending on the structure of the evolutionary fluctuations. In particular, distinct quasi-independent sectors may evolve instead of a single connected sector. A combinatorial process for generating new variations, involving for instance gene duplications, recombination events and/or horizontal transfers, has for instance been shown to produce modular organizations [11]. Such combinatorial variations may explain the modular organization of proteins into domains, which are subsequences of consecutive amino acids, but cannot easily account for the presence of multiple quasi-independent sectors distributed along the sequence of a single domain [4]. A scenario implicating modularly varying selective pressures provides an alternative explanation, as previously illustrated in a range of different models [12, 31, 37].

Consistently with these past works, we find that a modularly varying environment favors the emergence of two distinct sectors in an extension of our model where allostery involves two possible modulators. In this model,

the two modulators m_1, m_2 can bind at two distinct regulatory sites (Fig. 5), and selection is for preferential binding of the ligand ℓ in the presence of at least one of them (Material and methods). When both the sequences of m_1 and m_2 fluctuate in time, evolution stochastically generates one of two possible outcomes: systems with a single sector, as in Fig. 5A, or with two separate sectors, as in Fig. 5B. The probability to obtain two sectors depends on the structure of the fluctuations (besides the size of the structure): it is significantly larger when m_1 and m_2 change modularly, i.e., one at a time, compared to when they change non-modularly, i.e., simultaneously (Table in Fig. 5).

We note that, in contrast with previous models reporting similar effects [12, 31, 37], sparsity is not enforced in the definition of our fitness, but obtained as a result of evolution. We also find that a rugged fitness landscape is not necessary for modularity to emerge spontaneously [13]: in our model, solutions are indeed always accessible by hill-climbing with one-step mutations (Fig. 4B).

Discussion

Interpreting the information contained in the sequence of a protein does only require referring to its biophysical properties, but also to its evolutionary history. Our simple model of protein evolution thus demonstrates how a basic feature of proteins, the spatial organization of their residues least tolerant to mutations, may be controlled by past fluctuations of selective pressure or high mutation rates. Our conclusions are based on comparing scenarios that differ only in two evolutionary parameters, the period τ of environmental fluctuations and the mutation rate μ . Since a structural concentration of functional constraints arises only for some values of these parameters, it is clearly not a necessary consequence of the definition of our model.

We expect that comparable results hold for other systems where internal variables varying on a short time scale are similarly subject to short range interactions controlled by evolutionary and environmental variables varying on longer time scales. In less idealized systems, including natural proteins, several additional factors may, however, contribute to a concentration of functional constraints.

Irregular structures thus typically contain preferred allosteric paths that tend to reinforce the concentration of functional constraints: with no unique shortest path between its two interfaces, the cylindrical structure allowed us to illustrate the role of evolutionary factors with minimal contribution from structural heterogeneities. Our approach, however, extend to other geometries (Fig. S8).

Similarly, our results are robust to variations in the implementation of the evolutionary dynamics, but alter-

native choices may reduce or enhance sparsity; for instance, a multiplicative mutational process, which is biased towards vanishing couplings, generically favors sparsity over an additive process (Fig. S7) [38]. In our model, all coupling values are a priori equiprobable, showing not only that a mutational bias is not required, but also that sparsity of functional constraints, as reported by the fitness cost of mutations, does not equate sparsity of the underlying physical couplings (Fig. 2).

Sparsity may also be favored by factors limiting the efficiency of selection. The typically non-linear relationship between the biophysical properties of a protein and the reproductive rate (fitness) of organisms may thus make the contribution of all couplings unnecessary. Finite population size effects, which our genetic algorithm minimizes, also generically exclude a complete "optimization" of the couplings.

Our model represents an ideal case where, under constant environment, all the couplings may be equivalently involved in the function (the only a priori difference being between vertical and horizontal couplings). In the generic case where a uniform distribution of the couplings is intrinsically non-optimal, evolutionary fluctuations may, nevertheless, control the degree of concentration of functional constraints if they are sufficiently important.

Many extensions of our model are conceivable. Negative selection against undesired modulators and ligands may for example allow us to account for the specificity of the interactions. The assumption of a fixed geometry of interactions may also be relaxed to permit a joint treatment of folding and functional constraints, in line with previous studies based on similar simplified protein models [39, 40, 41]. Extending our model to account for structural changes and kinetic effects may thus contribute to rationalize the diversity of mechanisms that evolved to cause allostery [42].

Our model is not intended to account quantitatively for the features of natural proteins. Nevertheless, given the typical size $N \sim 10^2$ and mutations rates $\mu \sim 10^{-9}$ /bp/generation of current non-viral proteins, we may exclude a scenario based on high mutation rates for explaining the high tolerance of proteins to mutations. On the other hand, estimates of μP based on silent genomic variations within species give $\mu P \sim 10^{-1} - 10^{-3}$ for a range of organisms [43], where P represents an effective population size. This indicates that relevant time scales for a scenario based on fluctuating selective pressures are of the order of $\tau \sim (\mu P)^{-1} \sim 10 - 1000$ generations; these estimations are crude but lend weight to the plausibility of this scenario. Differences of variability in past selective pressures may thus cause different proteins to have fundamentally different architectures of functional constraints.

While our limited knowledge of past evolutionary history prevents us from testing quantitatively these ideas

with natural proteins, progress in the field of directed evolution [44, 45, 46] may soon offer us a platform to investigate them experimentally.

Materials and methods

Allosteric efficiency – A Gaussian spin-glass model is defined at inverse temperature β by the partition function [21]

$$Z(h|J) = \int \prod_i \frac{e^{-\sigma_i^2/2}}{\sqrt{2\pi}} d\sigma_i e^{-\beta H(\sigma|J,h)}, \quad (6)$$

where $H(\sigma|J,h) = -\frac{1}{2}\sigma^\top J\sigma - h^\top \sigma$ is the Hamiltonian of Eq. (4), with the geometry of the lattice defined by the non-zero elements of the matrix J_{ij} (with $J_{ii} = 0$ and $J_{ij} = J_{ji}$). The model is defined only at high temperature since the integral diverges for large β . If c denotes the maximal connectivity of the lattice, it is sufficient to assume that $|\beta h_i|, |\beta J_{ij}| < 1/c$, which, on a square lattice with $|h_i|, |J_{ij}| \leq 1$, we achieve by fixing $\beta = 0.1$. Under this assumption, the partition function is obtained by performing the Gaussian integration:

$$Z(h|J) = (\det(I - \beta J))^{-1/2} \exp\left(\frac{\beta^2}{2} h^\top (I - \beta J)^{-1} h\right), \quad (7)$$

where I represents the $M \times M$ identity matrix, M being the number of nodes in the lattice. The free energy $F(h|J) = -\beta^{-1} \ln Z(h|J)$ has hence the form

$$F(h|J) = -\frac{1}{2}\beta h^\top (I - \beta J)^{-1} h + F(0|J), \quad (8)$$

where $F(0|J)$ does not depend on h , leading to the following expression for $\phi(J|m, \ell)$, defined in Eq. (5):

$$\phi(J|m, \ell) = \beta m^\top [(I - \beta J)^{-1}]_{\text{reg,act}} \ell \quad (9)$$

where $[A]_{\text{reg,act}}$ denotes a sub-matrix of A_{ij} where i is restricted to $i \in \text{reg}$ and j to $j \in \text{act}$.

Evolutionary algorithm – Simulations were performed over 5.10^4 generations with populations of $P = 500$ systems consisting of 10×10 square lattices. At each generation, a system k with couplings J_{ij}^k is replicated n_k times based on the value of $\phi_k = \phi(J^k|m, \ell)$, following the sigma-scaling rule [28], $n_k = 1 + (\phi_k - \bar{\phi}) / (2\sigma_\phi^2)$, where $\bar{\phi}$ and σ_ϕ^2 are respectively the mean and variance of ϕ_k in the population. For each system, each coupling J_{ij} has then a probability μ to be mutated to a random value in $[-1, 1]$. See Suppl. Appendix for other replication and mutational rules.

Functionally significant couplings – To compare systems with different allosteric efficiencies, we define

the relative fitness cost of a mutation $J_{ij} \rightarrow J_{ij}^*$ as $\delta\phi_{ij}^*(J) \equiv (\phi(J) - \phi(J^*)) / \phi(J)$, where J^* differs from J by the value of J_{ij} . In Fig. 2B, we consider $\delta\phi_{ij}$, the highest fitness cost of a mutation at ij , estimated by comparing the fitness costs of $J_{ij}^* = 0, \pm 1$ and by retaining the highest one. A coupling is said to be functionally significant if $\delta\phi_{ij} > \epsilon$ with $\epsilon = 0.1$ (other choices of this cut-off yield similar results, see Fig. S1).

Sparsity – The sparsity S of a system is defined as the fraction of its couplings with $\delta\phi_{ij} < \epsilon$, using $\epsilon = 0.1$ (other choices of this cut-off yield similar results).

Sampling of systems with given fitness – To sample typical systems with a given value of fitness ϕ^0 , we implemented a standard Monte Carlo sampling algorithm with $|\phi(J|h) - \phi^0|$ as energy function.

Distance to alternative solutions – In Fig. 4B, the distance to an alternative selective pressure $h' \neq h$ is estimated as the number of steps necessary for a hill-climbing algorithm, whereby a single coupling J_{ij} can be changed at each step, to reach a fitness value $\phi(J|h')$ at least equivalent to the initial value $\phi(J|h)$.

Evolution of modularity – In a variant of our model, the regulatory site is split into two consecutive segments where two modulators m_1 and m_2 can bind, corresponding formally to $m = (m_1, m_2)$. Requiring the binding of a ligand ℓ to be allosterically regulated by the presence of any of the two modulators (non-exclusive OR) corresponds to selecting with a fitness $\phi = \min(\phi_1, \phi_2)$, where the allosteric efficiencies ϕ_1 and ϕ_2 are defined by Eq. (5) with, respectively, $m = (m_1, 0)$ and $m = (0, m_2)$. For the statistics shown in the table of Fig. 5, a system is considered as modular if removing the couplings below m_1 (by setting the couplings to 0) leads to a ϕ_2 within 80% of the original ϕ and removing those below m_2 to a ϕ_1 within 80% of the original ϕ (this definition is consistent with a classification based on visual inspection of networks as shown in Fig. 5; it is somewhat arbitrary but the trends shown in the table of Fig. 5 are not).

We thank A. Dawid, D. Hekstra, B. Houchmandzadeh, I. Junier, S. Leibler, C. Nizak, K. Reynolds, A. Raman and R. Ranganathan for discussions and comments. This work was supported by ANR grant CoevolInterProt.

-
- [1] J U Bowie, J F Reidhaar-Olson, W A Lim, and R Sauer. Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions. *Science*, 247:1306–1310, 1990.

- [2] S W Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286:295–299, 1999.
- [3] R N McLaughlin, Jr, F J Poelwijk, A Raman, W S Gosal, and R Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491:138–142, 2012.
- [4] N Halabi, O Rivoire, S Leibler, and R Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138:774–786, 2009.
- [5] F C Peterson, R R Penkert, B F Volkman, and K E Prehoda. Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. *Mol. Cell*, 13:665–676, 2004.
- [6] GM Süel, SW Lockless, MA Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10:59–69, 2003.
- [7] R G Smock, O Rivoire, W P Russ, J F Swain, S Leibler, R Ranganathan, and L M Gierasch. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol. Syst. Biol.*, 6:1–10, 2010.
- [8] K A Reynolds, R N McLaughlin, and R Ranganathan. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell*, 147:1564–1575, 2011.
- [9] L Hartwell, J Hopfield, S Leibler, and A Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.
- [10] GP Wagner, M Pavlicev, and J Cheverud. The road to modularity. *Nat. Rev. Gen.*, 8:921–931, 2007.
- [11] R V Solé and P Fernández. Modularity ”for free” in genome architecture? *arXiv preprint q-bio/0312032*, 2003.
- [12] N Kashtan and U Alon. Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci.*, 102:13773–13778, 2005.
- [13] J Sun and M W Deem. Spontaneous emergence of modularity in a model of evolving individuals. *Phys. Rev. Lett.*, 99:228107, 2007.
- [14] A Wagner. *Robustness and evolvability in living systems*. Princeton University Press, 2005.
- [15] G P Wagner and L Altenberg. Complex Adaptations and the Evolution of Evolvability. *Evolution*, 50:967–976, 1996.
- [16] L Ance and W Fontana. Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology*, pages 242–83, 2000.
- [17] J A Draghi, T L Parsons, G P Wagner, and J B Plotkin. Mutational robustness can facilitate adaptation. *Nature*, 463:353–355, 2010.
- [18] R D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Mol. Syst. Biol.*, 4:213, 2008.
- [19] H S Chan and E Bornberg-Bauer. Perspectives on protein evolution from simple exact models. *Applied Bioinformatics*, 1:121–144, 2002.
- [20] J D Bryngelson and P G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci.*, 84:7524–7528, 1987.
- [21] T H Berlin and M Kac. The spherical model of a ferromagnet. *Phys. Rev.*, 86:821, 1952.
- [22] I Bahar, A R Atilgan, and B Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2:173–181, 1997.
- [23] D M Taverna and R A Goldstein. Why are proteins so robust to site mutations? *J. Mol. Biol.*, 315:479–484, 2002.
- [24] G Tiana, B E Shakhnovich, N V Dokholyan, and E I Shakhnovich. Imprint of evolution on protein structures. *Proc. Natl. Acad. Sci.*, 101:2846–2851, 2004.
- [25] M Mézard, G Parisi, and M Virasoro. *Spin glass theory and beyond*. World Scientific, 1987.
- [26] J Monod, J P Changeux, and F Jacob. Allosteric proteins and cellular control systems. *J. Mol. Biol.*, 6:306–329, 1963.
- [27] P Leff. The two-state model of receptor activation. *Trends Pharm. Sci.*, 16:89–97, 1995.
- [28] M. Mitchell. *An Introduction to Genetics Algorithms*. MIT Press, 1999.
- [29] F Barahona. On the computational complexity of Ising spin glass models. *J. Phys. A*, 15:3241, 1982.
- [30] E van Nimwegen, J P Crutchfield, and M Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci.*, 96:9716–9720, 1999.
- [31] M Parter, N Kashtan, and U Alon. Facilitated variation: how evolution learns from past environments to generalize to new environments. *PLoS Comp. Biol.*, 4:e1000206, 2008.
- [32] M Eigen and P Schuster. *The hypercycle - a principle of natural self-organization*. Springer, 1979.
- [33] A D Keefe and J W Szostak. Functional proteins from a random-sequence library. *Nature*, 410:715–718, 2001.
- [34] L Ance Meyers, F D Ance, and M Lachmann. Evolution of genetic potential. *PLoS Comp. Biol.*, 1:236–243, 2005.
- [35] E Kussell, S Leibler, and A Grosberg. Polymer-population mapping and localization in the space of phenotypes. *Phys. Rev. Lett.*, 97:068101, 2006.
- [36] O Khersonsky and D S Tawfik. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Ann. Rev. Biochem.*, 79:471–505, 2010.
- [37] N Kashtan, A E Mayo, T Kalisky, and U Alon. An Analytically Solvable Model for Rapid Evolution of Modular Structure. *PLoS Comp. Biol.*, 5:e1000355, 2009.
- [38] T Friedlander, A E Mayo, T Tlusty, and U Alon. Mutation Rules and the Evolution of Sparseness and Modularity in Biological Systems. *PLoS ONE*, 8:e70444, 2013.
- [39] J D Hirst. The evolutionary landscape of functional model proteins. *Protein Eng.*, 12:721–726, 1999.
- [40] P D Williams, D D Pollock, and R A Goldstein. Evolution of functionality in lattice proteins. *J. Mol. Graphics and Modelling*, 19:150–156, 2001.
- [41] J D Bloom, C O Wilke, F H Arnold, and C Adami. Stability and the evolvability of function in a model protein. *Biophysical Journal*, 86:2758–2764, 2004.
- [42] H N Motlagh, J O Wrabl, J Li, and V J Hilser. The ensemble nature of allostery. *Nature*, 508:331–339, 2014.
- [43] M Lynch. The origins of eukaryotic gene structure. *Mol. Biol. Evol.*, 23:450–468, 2006.
- [44] P A Romero and F H Arnold. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, 10:866–876, 2009.
- [45] K M Esvelt, J C Carlson, and D R Liu. A system for the continuous directed evolution of biomolecules. *Nature*, 472:499–503, 2011.
- [46] A Fallah-Araghi, J-C Baret, M Ryckelynck, and A D Griffiths. A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab on a Chip*, 12:882–891, 2012.

SUPPLEMENTARY APPENDIX

Genetic algorithm – The results presented in the main text are obtained with the sigma-scaling procedure described in Materials and methods. This procedure ensures that the variance in the number of offsprings remains constant despite a decrease of variance in fitness. If taking the number of offsprings directly proportional to the fitness, the decrease over time of the fitness variance indeed renders selection ineffective (thus requiring longer simulations and/or larger populations). An alternative to sigma-scaling is an elite strategy whereby, at each generation, the top x % individuals with best fitness are duplicated while the bottom x % are eliminated. We verified that with $x = 20$, this strategy gives results equivalent to the sigma-scaling procedure.

Different modulators – In the main text, we present results when alternating between two opposite modulator sequences, $m^{(1)} = (+, \dots, +)$ and $m^{(2)} = (-, \dots, -)$. Any other choice of two opposite modulators with $m_i^{(1)} = -m_i^{(2)}$ for all $i \in \text{act}$ gives identical results as a consequence of the "gauge invariance": $\sigma_i \mapsto -\sigma_i \Leftrightarrow J_{ij} \mapsto -J_{ij} \forall j$. When alternating between two modulators $m^{(1)}$ and $m^{(2)}$ with $m_i^{(1)} = \pm 1$, $m_i^{(2)} = \pm 1$ and sequence similarity $s = \sum_i \delta(m_i^{(1)}, m_i^{(2)})$, the sparsity is commensurate with this measure of similarity (Fig. S6A), thus interpolating between the case $s = 10$, which is equivalent to a constant environment, and the case $s = 0$, which corresponds to opposite modulators.

In Fig. 4B, we consider systems that evolved under an environment fluctuating between two opposite modulators and, for a system at the end of a period of constant environment, represent the minimal number of mutations necessary to achieve an equivalent allosteric efficiency in the presence of the other modulator. Fig. S6B shows that similar results are obtained when considering a random modulator not previously encountered.

Alternative mutational processes – The results presented in the main text are obtained with memoryless mutations, consisting in drawing the new value of J_{ij} uniformly at random in $[-1, 1]$, independently of its previous value. Among other possible choices, we may consider: (i) discrete couplings, taken at random in a finite set of values, $\pm\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$; (ii) a sum-rule, where each mutation adds to the current value a normally distributed random variable: $J_{ij} \rightarrow J_{ij}^* = J_{ij} + \mathcal{N}(0, \sigma_s^2)$; (iii) a product-rule, where each mutation multiplies the current value by a Gaussian variable: $J_{ij} \rightarrow J_{ij}^* = J_{ij} \times \mathcal{N}(0, \sigma_p^2)$. We implemented these rules by mapping values $J_{ij}^* > 1$ to $J_{ij}^* = 1$ and values $J_{ij}^* < -1$ to $J_{ij}^* = -1$, to ensure that the couplings remain bounded. The results are presented

in Fig. S7, showing robustness of our conclusions with respect to the mutational process.

Alternative geometries – We presented our results with a two-dimensional square lattice but our model is equally solvable for any geometry. Fig. S8 thus shows results obtained with a $5 \times 5 \times 5$ three-dimensional regular cubic lattice with periodic boundary conditions along two dimensions and the regulatory and active sites defined on the faces associated with the third dimension (a three-dimensional generalization of the cylinder).

Coevolution – To analyze whether the functionally most significant couplings are subject to coevolution, we took a population obtained with $\tau = 200$, $\mu = 5.10^{-5}$, and used it as a common initial condition for 100 independent trajectory subject to the same τ, μ . After $3\tau = 600$ generations, we computed a matrix of covariance between couplings, $\mathcal{C}_{ij,kl} = |\langle J_{ij} J_{kl} \rangle - \langle J_{ij} \rangle \langle J_{kl} \rangle|$, where $\langle \dots \rangle$ denote an average over the different populations (the absolute value is taken to treat equivalently positive and negative covariations). We then performed a principal component analysis to identify the couplings that covary the most: Fig. 3B shows the matrix \mathcal{C} where the couplings are ordered based on their contribution to the principal eigenvector. The top positions defined by this principal component define a sector (Fig. 3C), which overlaps with the sector defined, as in Fig. 2B, based on the fitness cost of punctual mutations (Fig. 3A).

Robustness – The robustness of a system, shown in Figs. S9B-S10B, is defined as the fraction of mutations that do not cause a significant fitness cost. Formally,

$$R(J|h) = \langle \theta(\delta\phi_{ij}^* < 0.01) \rangle_{ij,*} \quad (10)$$

where $\theta(x) = 1$ if $x \geq 0$ and 0 otherwise, and where $\langle \dots \rangle_{ij,*}$ is an average over the pairs ij and over the possible values of J_{ij}^* .

Evolvability – The evolvability of a system, shown in Figs. S9C-S10C, is defined as the fraction of mutations that cause a significant advantage when the selective pressure changes. Formally,

$$E(J|h') = \langle \theta(\delta\phi_{ij}^* > 0.2) \rangle_{ij,*} \quad (11)$$

It differs from the definition of R by the field h' which is distinct from the field h in which the system most recently evolved. When considering an environment alternating periodically between two values $h^{(1)}$ and $h^{(2)}$, we thus take the systems at the end of a period of constant selective pressure under $h^{(1)}$ and define robustness as $R(J|h^{(1)})$ and evolvability as $E(J|h^{(2)})$.

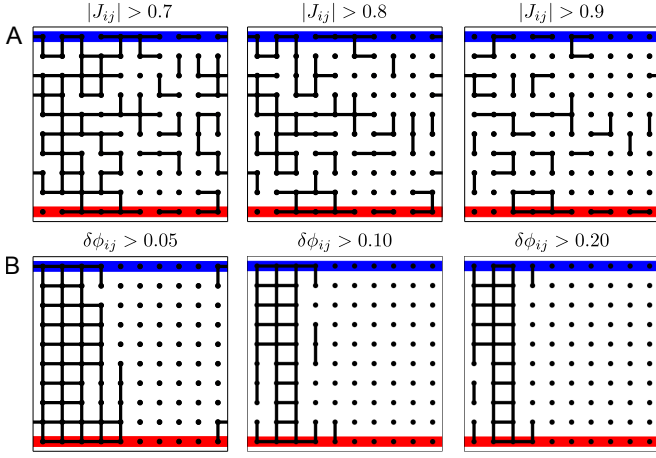


FIG. S1: For the system associated with $\tau = 200$ in Fig. 2, couplings $|J_{ij}|$ and functional constraints $\delta\phi_{ij}$ above different cut-offs (Fig. 2 corresponds to $|J_{ij}| > 0.8$ and $\delta\phi_{ij} > 0.1$).

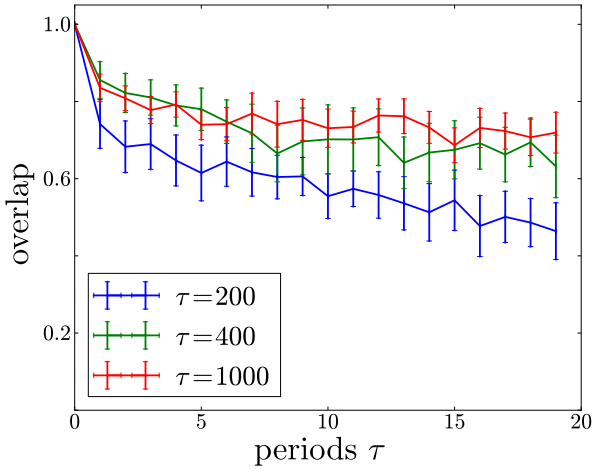


FIG. S2: Overlap of functionally significant couplings ($\delta\phi_{ij} > 0.1$) between a system at time t_0 and a system at time $t_0 + t$ along a same evolutionary trajectory as a function of time t , counted in number of periods τ (the error-bars are standard deviations over 100 simulations).

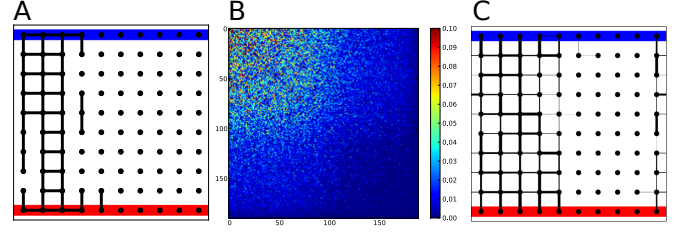


FIG. S3: Analysis of coevolution. **A.** Analogous to Fig. 2 with parameters $\tau = 200$ and $\mu = 5 \cdot 10^{-5}$. **B.** Matrix of covariations between couplings, obtained by comparing systems generated from independent trajectories originating from a common initial population; the positions are ordered by the principal eigenvector of the matrix. **C.** Mapping on the structure of the top couplings identified in B, showing a correspondence with the couplings identified in A based on the fitness cost of individual mutations.

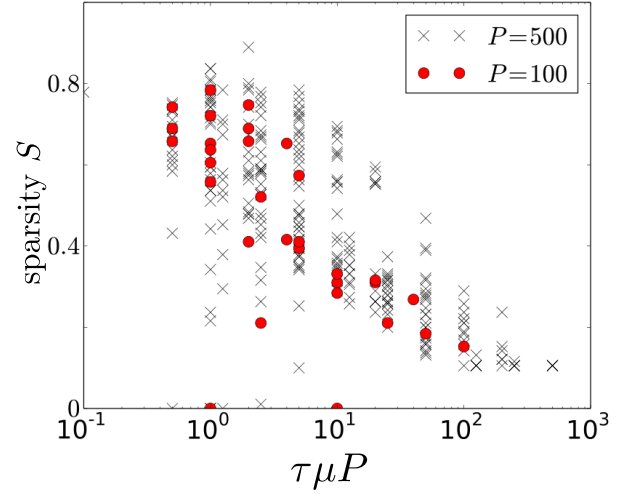


FIG. S4: Sparsity as a function of the scaling variable $\tau\mu P$ for systems obtained from evolutionary dynamics with different values of the period τ of environmental changes and mutation rate μ (in the range of values used for Fig. 3) and for two population sizes, $P = 100, 500$.

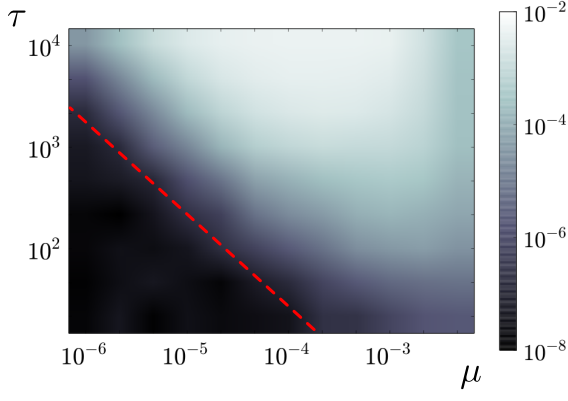


FIG. S5: Fitness of the population as a function of μ and τ , the axis and the red line are the same as in Fig. 3. The red line ($\phi = 10^{-7}$) corresponds to the typical maximal value of allosteric efficiency in populations of $P = 500$ random systems. Below this line, the populations may be considered as non-adapted.

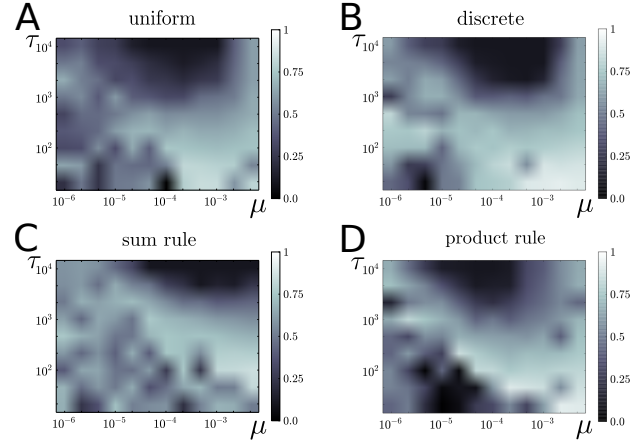


FIG. S7: Sparsity as a function of mutation rate μ and time scale τ of environmental changes for systems that evolved subject to different mutational processes: **A.** As in Fig. 3, the J_{ij} are mutated to a random value uniformly distributed in $[-1, 1]$, independently of their previous value. **B.** The J_{ij} are drawn from a finite set of discrete values. **C.** Sum-rule with variance $\sigma_s^2 = 0.2$. **D.** Product-rule with variance $\sigma_p^2 = 1.6$. In each case and as in Fig. 3, the left bottom corner corresponds to non-adapted populations.

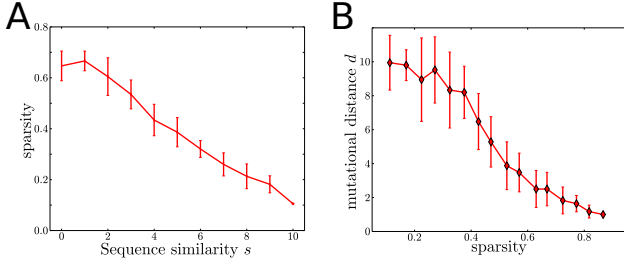


FIG. S6: **A.** Sparsity as a function of the sequence similarity s between the two alternative modulator sequences, for $\mu = 10^{-5}$ and $\tau = 100$ (mean and standard deviation over 10 simulations). **B.** Minimal number of point mutations necessary to adapt to a new random modulator as a function of sparsity. The simulations are obtained with different values of $\tau \in [10, \dots, 5000]$ and $\mu \in [5 \cdot 10^{-3}, \dots, 2 \cdot 10^{-6}]$, when they correspond to adapted populations ($\phi > 10^{-7}$); the results are averages over 5 random modulators.

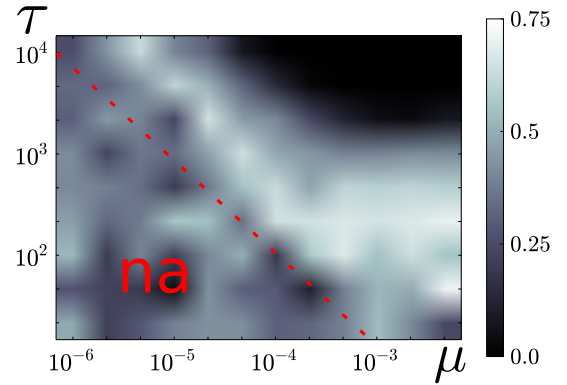


FIG. S8: Sparsity of evolved systems as a function of μ and τ for systems defined on a three-dimensional $5 \times 5 \times 5$ square lattice with periodic conditions along two dimensions and regulatory and active sites at the two open boundaries of the third. These results generalize those of Fig. 3 to a three-dimensional system.

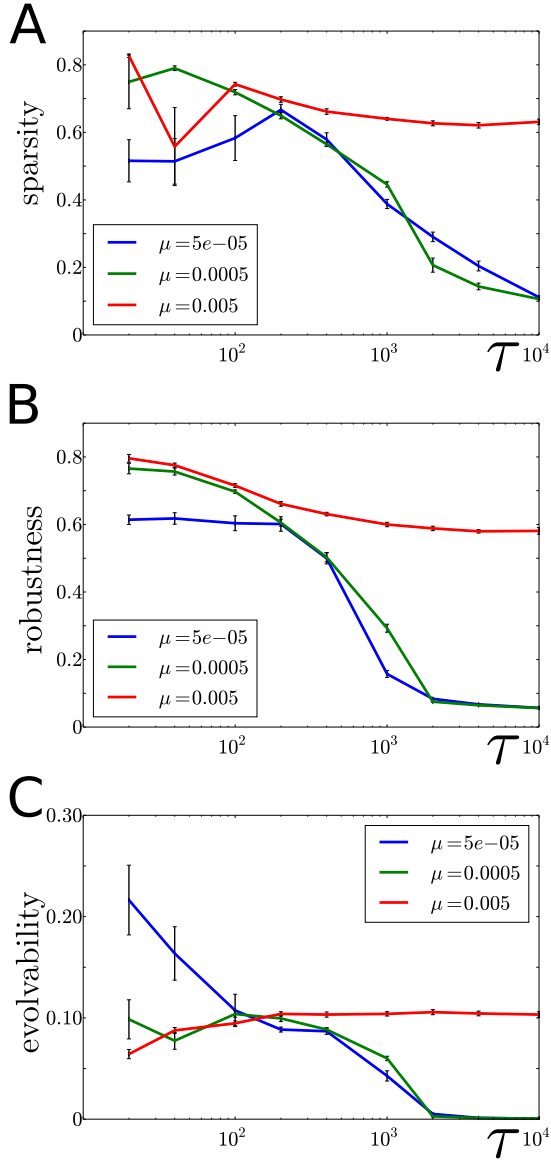


FIG. S9: Sparsity, robustness and evolvability of evolved systems as a function of the period τ of environmental changes. The error-bars are standard deviations over 10 simulations.

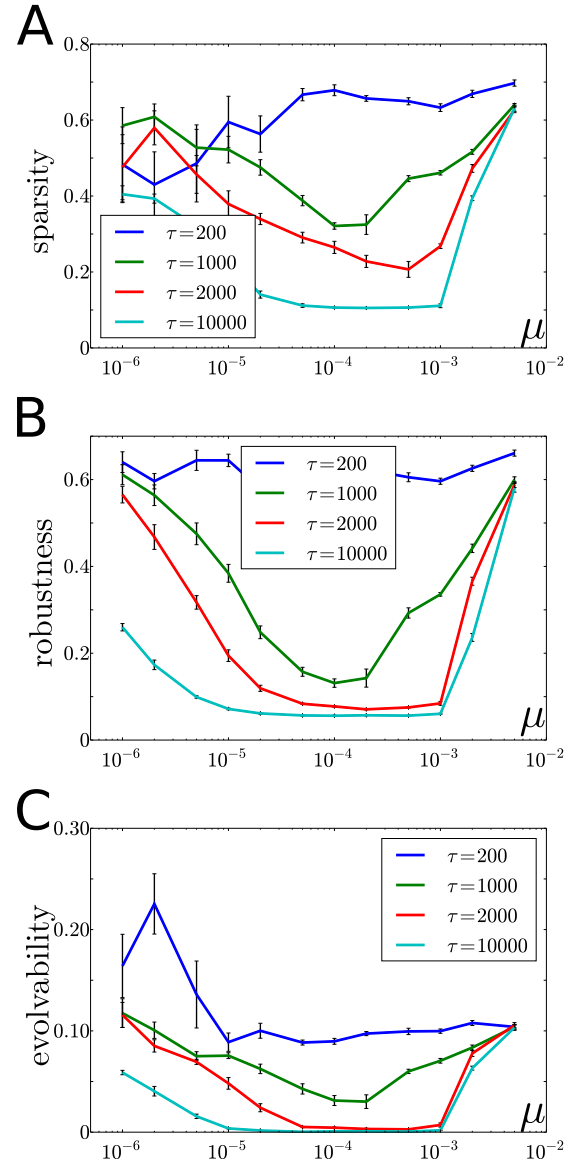


FIG. S10: Sparsity, robustness and evolvability of evolved systems as a function of the mutation rate μ . The error-bars are standard deviations over 10 simulations.